22S:105
Statistical Methods and Computing

Review of statistical significance, one-sample z tests, and z confidence intervals for a population mean; Power and sample size

Lecture 15
Mar. 25, 2031

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

What is the **power** of this test?

This depends on the particular alternative mean that we are interested in.

For computing power, we have to rewrite our alternative hypothesis to specify a **single** value of interest.

- This value is determined by **subject matter** considerations, not statistical considerations.
- Suppose you considered my husband's true resting pulse rate would have to be at least 5 beats higher than 45 in order for the difference to be worth caring about. Then:

$$H_0 : \mu = \mu_0 = 45$$
$$H_a : \mu = \mu_1 = 50$$

If H_a is true, then the sampling distribution of $\bar{x}$ is normal with mean $\mu_1 = 50$ oz and standard deviation $\frac{\sigma}{\sqrt{n}} = 1.79$ oz.

$$\begin{aligned} power &= P(reject\ H_0 \mid H_0\ false) \\ &= P(reject\ H_0 \mid H_a\ true) \\ &= P(\bar{X} \geq 47.9 \mid \mu = 50) \\ &= P(Z \geq \frac{47.9 - 50}{1.79}) \\ &= P(Z \geq -1.173) \\ &= 1 - .121 \\ &= 0.879 \end{aligned}$$

Therefore, there is about an 88% chance of correctly rejecting H_0 using a .05 significance level and a sample size of 5.

Statistical significance

- an observed result in a study that is very unlikely to have occurred by chance is said to be *statistically significant*
- When we specify the *significance level* α as part of the planning of a study, we are defining how much evidence against the null hypothesis we will require.
  - how large a probability of rejecting H_0 if it is actually true that we can tolerate
  - how small the p-value will have to be in order for us to reject H_0
  - e.g., if we choose $\alpha = .01$ we are requiring that the data give evidence against H_0 so strong that it would happen no more than 1% of the time if H_0 is true
    - i.e. no more than 1 time in 100 samples

Test statistics

- Hypothesis tests generally are based on a *statistic* that estimates the unknown parameter that appears in the hypotheses.
  - When the parameter of interest is the population mean μ, the appropriate statistic is the sample mean $\bar{x}$.
  - When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0.
  - Values of the estimate far from the parameter value in H_0 give evidence against H_0. H_a determines which direction(s) count against H_0.

  - In order to quantify how likely or unlikely the value of the statistic would be if H_0 were true, we need to convert the statistic to a *test statistic* – a random variable with a distribution that we know.
    * If the parameter of interest is a population mean μ, and we assume that the population distribution is normal with a known σ, then the appropriate test statistic is a z statistic.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

More on statistical significance

- If, after we have conducted the study and carried out the hypothesis test, the p-value is as small as or smaller than α, the result is said to be *statistically significant at level* α.
- Statistical significance is **not** the same as practical significance.
  - Example: Suppose you took 625 measurements of my husband's pulse and got an $\bar{x}$ of 45.5.
    The z statistic would be

$$z = \frac{45.5 - 45}{4/25} = 3.13$$

    and the p-value for a 1-sided test would be 0.0009.
    But who would care when the actual difference was so small?

- Significance levels and p-values are meaningless for poorly designed studies.

- In order for the results of a hypothesis test to be meaningful, the investigators must decide *in advance* exactly what hypotheses they wish to test.
  - Suppose we were aware of 100 different variables that characterize UI students.
  - We might be tempted to carry out 100 different hypothesis tests to determine whether mean GPA was different in subgroups defined by different values of these variables.
  - If we did each of 100 hypothesis tests at significance level $\alpha = .05$, we would expect to reject the null hypothesis on about 5 of them by pure random chance even if none of the variables had any relationship to GPA.

  - Exploratory analysis to seek possibly relevant variables is fine and necessary, but hypothesis tests should not be done and statistical significance should not be assessed.
  - If some possible variables of interest are identified in exploratory analysis, then it would be appropriate to do a hypothesis test later *with different data* to determine whether the association is real.

Another example of power for one-sided hypothesis tests

Recall

- $\alpha = P(reject\ H_0 \mid H_0\ true)$
- $\beta = P(fail\ to\ reject\ H_0 \mid H_0\ false)$
- **power** of the test = $1 - \beta = P(reject\ H_0 \mid H_0\ false)$

Example:

We wish to test the hypothesis that mothers under 16 years of age deliver babies whose birthweights are in some sense lower than "normal." We know from large nationwide surveys based on millions of deliveries that the mean birthweight in the US is 120 oz with a standard deviation of 25 oz.

Our hypotheses are:

$$H_0 : \mu = 120$$
$$H_a : \mu < 120$$

We will

- choose the significance level $\alpha = .05$ for our test.
- assume that the population standard deviation of birthweights to women under 16 years of age is the same as in the general population of women; that is, $\sigma = 25$ oz.
- study a simple random sample of 100 mothers under age 16

What value of $\bar{x}$ would it take for us to reject H_0 and conclude that the population mean of birthweights to women under 16 years of age is lower than 120 oz?

We know that, if H_0 is true, then the sampling distribution of $\bar{x}$ is normal with

- mean 120 oz
- standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{25}{10} = 2.5$ oz

We need to find the value of $\bar{x}$ that would cut off the *lower* .05 of the area under this normal curve.

For a standard normal curve, -1.645 cuts off the lower .05 of the area.

To convert to the sampling distribution of $\bar{x}$

$$\bar{x} = \mu_0 + z^* \frac{\sigma}{\sqrt{n}}$$

$$\begin{aligned} &= 120 - 1.645(2.5) \\ &= 115.89 \end{aligned}$$

So we would reject H_0 if we obtained a sample mean $\bar{x} \leq 115.89$ oz.

What is the **power** of this test?

This depends on the particular alternative mean that we are interested in.

For computing power, we have to rewrite our alternative hypothesis to specify a single value of interest.

$$\begin{aligned} H_0 &: \mu = \mu_0 = 120 \\ H_a &: \mu = \mu_1 = 115 \end{aligned}$$

If H_a is true, then the sampling distribution of $\bar{x}$ is normal with mean $\mu_1 = 115$ oz and standard deviation $\frac{\sigma}{\sqrt{n}} = 2.5$ oz.

$$\begin{aligned} power &= P(reject\ H_0 \mid H_0\ false) \\ &= P(reject\ H_0 \mid H_a\ true) \\ &= P(\bar{X} \leq 115.89 \mid \mu = 115) \\ &= P(Z \leq \frac{115.89 - 115}{2.5}) \\ &= P(Z \leq .356) \\ &= 0.641 \end{aligned}$$

Therefore, there is about a 64% chance of detecting a significant difference using a .05 significance level and a sample size of 100.

Factors that affect power

- If the *significance level* α decreases, the power decreases.
- If the alternative mean is shifted farther away from the null mean ($|\mu_0 - \mu_1|$ increases), the power increases.
- If the standard deviation σ of an individual observation increases, power decreases.
- If the sample size n increases, the power increases.

Sample size

If we are planning a study that involves a hypothesis test, we should

- decide on the significance level α
- decide the desired power
- decide on the specific alternative hypothesis of interest
- compute the number of items we will need in our sample

Example: Suppose we wanted to have 80% power instead of 64% power in the birthweights problem.

- $\alpha = .05$
- power is .80
- $\mu_0 = 120$ and $\mu_1 = 115$
- We need to solve for n.

Since $\alpha = .05$, we again note that H_0 would be rejected for $z \leq -1.645$). Substituting $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ for z, we get

$$-1.645 = \frac{\bar{x} - 120}{25/\sqrt{n}}$$

$$\bar{x} = 120 - 1.645\frac{25}{\sqrt{n}}$$

So we would reject H_0 if $\bar{x}$ is less than 120 - 1.645 $\frac{25}{\sqrt{n}}$.

Now consider the desired power of the test. If the true mean birth weight were $\mu_1 = 115$, we would want to reject H_0 with probability 0.80, or to fail to reject it with probability $\beta = 0.20$.

The value of z that corresponds to $\beta = 0.20$ is z = 0.84. Therefore

$$0.84 = \frac{\bar{x} - 115}{25/\sqrt{n}}$$

$$\bar{x} = 115 + 0.84\frac{25}{\sqrt{n}}$$

Setting the two expressions for the sample mean $\bar{x}$ equal to each other gives:

$$120 - 1.645\frac{25}{\sqrt{n}} = 115 + 0.84\frac{25}{\sqrt{n}}$$

or

$$\sqrt{n}(120 - 115) = (0.84 + 1.645)(25)$$

and

$$\begin{aligned} n &= \left[\frac{(0.84 + 1.645)(25)}{120 - 115}\right]^2 \\ &= 154.38 \end{aligned}$$

We would need a sample size of 155 to have an 80% chance of detecting a significant difference at the .05 level if the althernative mean is 115 oz.

The generic formula for sample size estimation for a hypothesis regarding the population mean μ of a normal distribution

- one-sided alternative
- σ assumed known
- using normal table that gives areas to the *left* of a cutoff

$$n = \frac{\sigma^2(z^*_{(1-\beta)} + z^*_{(1-\alpha)})^2}{(\mu_0 - \mu_1)^2}$$

Factors affecting the sample size

- The sample size increases as σ increases.
- The sample size increases as the significance level is made smaller (α decreases)
- The sample size increases as the required power increases
- The sample size decreases as the distance between the null and alternative means increases